Displays for Statistics 5303

Lecture 6

September 16, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)
612-625-1024 (Minneapolis)

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303





## More about ANOVA

An F-test in an analysis of variance is actually a test for a specific comparison of two two hypothesis, each specifying a model.

In the one-way ANOVA,

$H_0$: $\mu_1 = \mu_2 = \ldots = \mu_g = \mu^*$

or $\alpha_1 = \alpha_2 = \ldots \alpha_g = 0$

**Model** is $y_{ij} = \mu^* + \varepsilon_{ij}$

$H_a$: At least two $\mu_i$'s differ

or at least two $\alpha_i$'s differ

**Model** is $y_{ij} = \mu_i + \varepsilon_{ij} = \mu^* + \alpha_i + \varepsilon_{ij}$

As a model, $H_a$ is sometimes called the *unrestricted model* or the *full model*.

Suppose you knew $H_0$ were true.

- Your best estimate of $\mu^*$ would be $\overline{y}_{\bullet\bullet}$.
- The residuals would be $y_{ij} - \overline{y}_{\bullet\bullet}$
- The residual SS would be $SSR_0 = SS_T = \sum\sum(y_{ij} - \overline{y}_{\bullet\bullet})^2$.

In the unrestricted case ($H_A$),

- Your best estimates of $\mu_1, \ldots, \mu_g$ are $\overline{y}_{1\bullet}, \ldots, \overline{y}_{g\bullet}$
- The residuals would be $y_{ij} - \overline{y}_{i\bullet}$
- The residual SS would be $SSR_A = SS_E = \sum_i\sum_j(y_{ij} - \overline{y}_{i\bullet})^2$.

Thus $SS_{trt} = SS_T - SS_E = SSR_0 - SSR_A$ is the *reduction in the residual SS* you can achieve if you leave $H_0$ in favor of $H_a$ and

$$F = \{SS_{trt}/(g-1)\}/\{SS_E/(N-g)\}$$

is a way to see if this reduction is large enough to be significant.





This is a general principle used in ANOVA and regression:

$$F = \frac{(SSR_0 - SSR_A)/(df_0 - df_A)}{SSR_A/df_A}$$

Where

$df_A = N - n_A$, $n_A$ = #parameters for $H_a$)
$= N - g$ (in this case)

and

$df_0 = N - n_0$, $n_0$ = #parameters for $H_0$)
$= N - 1$ (in this case)

$df_{trt} = df_0 - df_A = N - 1 - (N - g) = g - 1$
$df_{error} = df_A = N - g$

**Comment** The ratio $SS_{trt}/SS_T$ is the proportion of the total variation that can be "explained" by differential effects of treatments. It is the direct analogue of the coefficient of determination (multiple $R^2$) in regression.

Why is this effective? It all depends on the **expectations of mean squares** (MS) in the ANOVA.

Suppose $H_0$ is true. Then

$$E[SSR_0] = df_0\sigma^2 = (N - 1)\sigma^2$$

and

$$E[SSR_A] = \sum(n_i-1)\sigma^2 = df_A\sigma^2 = (N-g)\sigma^2$$

Therefore

$$E[SS_{trt}] = E[SSR_0] - E[SSR_A]$$

$$(N - 1)\sigma^2 - (N - g)\sigma^2 = (g - 1)\sigma^2 = df_{trt}\sigma^2$$

$$E[SS_{error}] = E[SSR_1] = (N-g)\sigma^2 = df_{error}\sigma^2$$

Since mean squares are SS/df,

$$E[MS_{trt}] = E[SS_{trt}/(g-1)] = \sigma^2$$

$$E[MS_{error}] = E[SS_{error}/(N-g)] = \sigma^2$$

**Conclusion**: When $H_0$ is true, the expectation of both the numerator and denominator of $F = MS_{trt}/MS_{error}$ are the same. The median of F is close to 1.





When $H_0$ is not *true*, it is still true that

$$E[MS_{error}] = E[SS_{error}/(N-g)] = \sigma^2$$

Now, however,

$$E[MS_{trt}] = \sigma^2 + \tau^2/(g-1) > \sigma^2$$

where

$$\tau^2 \equiv \sum_{1\leq i\leq g} n_i(\mu_i - \tilde{\mu})^2, \quad \tilde{\mu} = \sum n_i\mu_i/N.$$

Note that $\tau^2 = 0$ when $H_0$ is true.

So violation of $H_0$ *increases* $E[MS_{trt}]$ and hence E[F], and makes it more probable you will reject $H_0$.

If you use the parametrization which sets $\mu^* = \tilde{\mu} = \sum n_i\mu_i/N$ and $\alpha_i = \mu_i - \mu^*$,

$$\tau^2 \equiv \sum_{1\leq i\leq g} n_i\alpha_i^2$$

This formula is *not* correct for other choices for $\mu^*$. In particular it is not true for $\mu^* = \overline{\mu} = \sum\mu_i/g$, $\alpha_i = \mu_i - \overline{\mu}$ (unless the sample sizes are equal).

To **summarize**:

Testing an ANOVA hypothesis is equivalent to a comparison of two models

- a null model

and

- a more general alternative model.

Your conclusion in the test is effectively a *selection* of one or the other model as most appropriate.

In more complex ANOVA's you may be selecting among more than two models.





## Contrasts

A contrast is a formula which compares two or more treatment means or effects in a way that <u>doesn't depend on the overall level $\mu^*$</u>.

**Examples**:

- $\mu_1 - \mu_3 = (\mu^* + \alpha_1) - (\mu^* + \alpha_3) = \alpha_1 - \alpha_3$
- $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4 + \mu_5)/3$
  $= (\mu^*+\alpha_1 + \mu^*+\alpha_2)/2$
  $\quad - (\mu^*+\alpha_3 + \mu^*+\alpha_4 + \mu^*+\alpha_5)/3$
  $= (\alpha_1 + \alpha_2)/2 - (\alpha_3 + \alpha_4 + \alpha_5)/3$

This compares the average of the first 2 means or effects with the average of the last 3.

## Formal definition

A *contrast* is a linear combination of $\mu$'s
$w(\{\mu_i\}) \equiv \sum_i w_i \mu_i$, with $\sum_i w_i = 0$

Because $\sum_i w_i = 0$, $w(\{\mu_i\})$ doesn't depend on $\mu^*$:

$$\begin{aligned}\sum_i w_i \mu_i &= \sum_i w_i(\mu^* + \alpha_i)\\ &= (\sum_i w_i)\mu^* + \sum_i w_i \alpha_i = \sum_i w_i \alpha_i\\ &= 0 \times \mu^* + \sum_i w_i \alpha_i = \sum_i w_i \alpha_i\\ &= w(\{\alpha_i\})\end{aligned}$$

Since $\sum_i w_i \alpha_i$ doesn't depend on $\mu^*$ this satisfies the informal definition of a contrast given before.

The weights $\{w_i\}$ themselves are also often referred to as a *contrast*.

An *observed contrast* is
$w(\{\overline{y}_{i\bullet}\}) = \sum_i w_i \overline{y}_{i\bullet} = \sum_i w_i \hat{\alpha}_i = w(\{\hat{\alpha}_i\})$





You may sometime calculate several contrasts as part of your analysis.

What you use depends on the *questions of interest* to the researcher, not on some statistical magic.

If you are just providing statistical advice, you need to find out what questions need answers.

More on the example:

```
Cmd> anova("logy=treat",fstat:T)
Model used is logy = treat
WARNING: summaries are sequential
              DF         SS          MS            F       P-value
CONSTANT       1     79.425      79.425   8653.95365      < 1e-08
treat          4     3.5376     0.88441     96.36296      < 1e-08
ERROR1        32    0.29369   0.0091779
```

Compute $\hat{\mu}_i = \overline{y}_{i\bullet}$ and $\hat{\alpha}_i = \overline{y}_{i\bullet} - \sum \overline{y}_{i\bullet}/g$

```
Cmd> muhats <- tabs(logy,treat,mean:T);muhats
(1)      1.9325       1.6287       1.3775       1.1943       1.0567

Cmd> alphahats <- muhats - sum(muhats)/5; alphahats
(1)     0.49456      0.19081     -0.06044     -0.24365     -0.38127
```

MacAnova function **coefs()** computes $\hat{\alpha}_i$'s

```
Cmd> coefs(treat) # or coefs("treat") or coefs(2)
(1)     0.49456      0.19081     -0.06044     -0.24365     -0.38127
```

**coefs(2)** would also work too because **treat** is line 2 in **anova()** output.

Enter weights and compute contrast two ways.

```
Cmd> w <- vector(vector(1,1)/2,-vector(1,1,1)/3); w
(1)        0.5        0.5   -0.33333   -0.33333   -0.33333

Cmd> vector(sum(w),sum(w*muhats),sum(w*alphahats))
(1)  1.1102e-16    0.57114    0.57114
```

MacAnova function **contrast()** makes it easy to compute contrasts.

```
Cmd> stuff <- contrast(treat, w); stuff
component: estimate
(1)    0.57114      Same as just computed
component: ss
(1)     2.9446
component: se
(1)   0.031886
```

The result (output) from **contrast()** is a *structure* with three *components*:

- The **estimate** component is the value of the contrast $\sum w_i\hat{\alpha}_i$.
- The **se** component is its estimated standard error. You can compute a t-statistic to test the null hypothesis that the $\sum w_i\alpha_i = 0$
- The **ss** component is an SS associated with the contrast.





Using **estimate** and **se** to compute a t-statistic to test $H_0$: $\sum_i w_i\alpha_i = 0$:

```
Cmd> tstat <- stuff$estimate/stuff$se

Cmd> vector(tstat,twotailt(tstat,DF[3]))
(1)     17.912  3.0663e-18    t-statistic and P-value
```

When $H_0$ is true, t has Student's t-distribution on $df_{error}$ = N - g d.f.

**stuff$estimate** is one way to extract a component from a structure. Since this is the first component, another way is **stuff[1]** and t is **stuff[1]/stuff[3]**.

The **ss** component (**stuff$ss** or **stuff[2]**) is MSE×estimate$^2$/se$^2$

```
Cmd> mse <- SS[3]/DF[3]; mse # MS in ERROR1 row of ANOVA
     ERROR1
  0.0091779

Cmd> mse*(stuff$estimate/stuff$se)^2
(1)     2.9446
```

**stuff$ss/mse** is the same as $t^2$:

```
Cmd> vector(stuff$ss/mse, tstat^2)
(1)     320.83     320.83
```

## Common Contrasts

- Pairwise contrasts

  $\mu_i - \mu_j = \alpha_i - \alpha_j$ Compare two groups

  $\{w_i\} = \{0,\ldots,1,0,\ldots,-1,0,\ldots\}$

For g groups, there are g(g-1)/2 essentially different pairwise contrasts:

```
Cmd> print(pairwise_wts,format:"4.0f")
pairwise_wts: Each columns is a set of contrast weights
(1,1)    1    1    1    1    0    0    0    0    0    0
(2,1)   -1    0    0    0    1    1    1    0    0    0
(3,1)    0   -1    0    0   -1    0    0    1    1    0
(4,1)    0    0   -1    0    0   -1    0   -1    0    1
(5,1)    0    0    0   -1    0    0   -1    0   -1   -1
```

The following computes the contrast and t-statistic for the 5×4/2 = 10 pairwise contrasts.

```
Cmd> for(i,1,10){
        stuff <- contrast(treat,pairwise_wts[,i])
        print(paste("W =", pairwise_wts[,i],",estimate =",\
           stuff$estimate,", t-statistic =",\
           stuff$estimate/stuff$se))
     }
W = 1 -1 0 0 0 ,estimate = 0.30375 , t-statistic = 6.3413
W = 1 0 -1 0 0 ,estimate = 0.555 , t-statistic = 11.586
W = 1 0 0 -1 0 ,estimate = 0.73821 , t-statistic = 14.889
W = 1 0 0 0 -1 ,estimate = 0.87583 , t-statistic = 16.928
W = 0 1 -1 0 0 ,estimate = 0.25125 , t-statistic = 5.2452
W = 0 1 0 -1 0 ,estimate = 0.43446 , t-statistic = 8.7626
W = 0 1 0 0 -1 ,estimate = 0.57208 , t-statistic = 11.057
W = 0 0 1 -1 0 ,estimate = 0.18321 , t-statistic = 3.6952
W = 0 0 1 0 -1 ,estimate = 0.32083 , t-statistic = 6.201
W = 0 0 0 1 -1 ,estimate = 0.13762 , t-statistic = 2.582
```





- **Comparison with control**

Say treatment 1 is a control.

An obvious idea is to compare the mean or effect of the control with the average mean or effect of all the non-controls:.

$$\mu_1 - (\mu_2 + \mu_3 + \ldots + \mu_g)/(g-1)$$
$$= \alpha_1 - (\alpha_2 + \alpha_3 + \ldots + \alpha_g)/(g-1)$$

Contrast coefficients are

$$\{w_i\} = \{1, -1/(g-1), \ldots, -1/(g-1)\}$$

Of course individual pairwise comparisons with control $\alpha_i - \alpha_1$, i = 2, ..., g would probably of interest too.

Multiplying this by g - 1, an equivalent contrast is

$$(g-1)\mu_1 - \mu_2 - \mu_3 - \ldots - \mu_g$$

with integer coefficients {g-1,-1,...,-1}. Before computers were common, this made calculations easier.

- **Factorial treatments**

When there are two factors, A and B, each at two levels, there are 4 treatments with means $\mu_{11}$, $\mu_{12}$, $\mu_{21}$, $\mu_{22}$. These can be displayed in a 2 by 2 table

| | $B_1$ | $B_2$ |
|---|---|---|
| $A_1$ | $\mu_{11}$ | $\mu_{12}$ |
| $A_2$ | $\mu_{21}$ | $\mu_{22}$ |

Natural contrasts would be

- Average of row 1 vs average of row 2:

$$(\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$$

This measures the effect of factor A, ignoring factor B (main effect of A).

- Average of col. 1 vs average of col. 2:

$$(\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2$$

This measures the effect of factor B, ignoring factor A (main effect of B).





- Difference between effects of A for the two levels of B

$$(\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$$

This is algebraiclly the same as the difference between effects of B for the two levels of A

$$(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$$

When this contrast is not zero, it means the effect of A depends on the level of B (or the effect of B depends on the level of A).

When this occurs, we say there is *interaction* between factors A and B. So this is an *interaction* contrast.

Suppose the treatments are determined by a quantitative variable x with levels $x_1, x_2, \ldots, x_g$, say. Then, if you fit a straight line, the least squares estimate of the slope is

$$b = \sum n_i(x_i - \overline{x})\overline{y}_{i\bullet}/\sum n_i(x_i - \overline{x})^2,\ \overline{x} = \sum n_i x_i/N$$

When the sample sizes are equal, you can omit the $n_i$.

This is a contrast with weights

$$w_i = n_i(x_i - \overline{x})/\sum_i n_i(x_i - \overline{x})^2$$

which do satisfy $\sum w_i = 0$, because $\sum_i n_i(x_i - \overline{x}) = 0$.

It will be large when there is a high degree of linear dependence of the means on x.

This is a *linear* contrast because it focuses on the strength of a straight line relationship between $\mu_i$ or $\alpha_i$ and $x_i$.